

## A DATA SWITCH AND A METHOD FOR BROADCAST PACKET QUEUE ESTIMATION

Field of the invention

The present invention relates to a data switch and to a method of operating it.

Background of Invention

5 One of the types of data packets which Ethernet switches are required to transmit are broadcast packets, i.e. packets which are to be transmitted from one of the ingress ports to all of the egress ports, except the egress port corresponding to the ingress port ("source port") from which the broadcast packet arrived. Shared memory output queue Ethernet switches cannot  
10 sustain excessive levels of broadcast packets, because the memory requirements increase linearly with the percentage of broadcasts in a traffic stream. This means that there is a need to limit the number of broadcasts in the system.

In the case that it is identified that the number of broadcast packets is  
15 excessive, it is known to delete selected ones of the broadcast packets, e.g. selectively based on a parameter in the header of the packet defining the importance of the packet. This is referred to as "broadcast storm control" (BSC).

Conventional methods to identify excessive amounts of broadcast packet  
20 traffic operate by counting the number of broadcasts per unit time. Once this value rises above a predefined level, BSC is turned on. When the figure drops below the predetermined level (e.g. by a certain amount, so that there is a hysteresis), BSC is turned off. This method suffers from the problem that it requires a counter for explicitly counting the broadcast packets. Additionally,  
25 since the count must be worked out per unit time, a timer is required, e.g. to decrement the counter every timer interval.

### Summary of the Invention

The present invention proposes a new and useful manner of determining excess levels of broadcast packets, in particular so that BSC can be carried out.

- 5 In general terms the invention proposes that the length of the respective queues at the ingress ports is measured, and the level of broadcast packets is estimated, or in some circumstances exactly determined, based on these lengths. The method is motivated by the observation that a broadcast packet takes longer than a normal packet to pass through the switch, and therefore  
10 causes the length of the queue to grow. In wirespeed unicast systems with one-to-one traffic flow, broadcast packets are in fact the only types of packets which can cause the ingress queues to lengthen.

- From the level of the broadcast packets, a determination is made of whether or not the level is excessive, and in this case BSC can be carried out, for  
15 example according to the conventional methods described above. For example, BSC can be carried out whenever the system determines that the length of any of the queues rises above a predetermined level, since the length of that queue provides a measure of the frequency of arrival of broadcast packets (at the corresponding ingress port).

### 20 Brief Description of The Figures

Preferred features of the invention will now be described, for the sake of illustration only, with reference to Fig. 1, which shows schematically a switch according to the invention.

### 25 Detailed Description of the embodiments

Referring to Fig. 1, a Ethernet switch which is an embodiment of the invention is shown. According to conventional structures, the Ethernet switch has a number of ingress ports  $n$  and a corresponding number  $n$  of egress ports. Data packets arrive at the ingress ports for transmission across a switching  
5 fabric to the egress ports.

The Ethernet switch has a packet resolution module 1 including a respective ingress queue 3 for each ingress port. The ingress queues are marked from  $Rx\#0$  up to  $Rx\#n-1$ . The packet resolution module 1 determines a destination list for each packet arriving at a certain ingress port (i.e. a list of the egress  
10 ports to which it should be transmitted), and stores this information in the corresponding queue. The destination list for a typical packet is labelled 4 in Fig. 1, and includes for each of the  $n$  destinations either an indication that the packet is to be sent there (marked in destination list 4 as a black square), or that it is not (marked as a 0). The destination list 4 shown in Fig. 1 is for a  
15 broadcast packet having ingress 1 as the source port, so that it is 0 for destination 1, and a black square for all other destinations.

The Ethernet switch further includes a queue management module 5 having a scheduler 7 and a respective egress queue 9 for each of the  $n$  egress ports.  
20 The egress queues are marked from  $Tx\#0$  up to  $Tx\#n-1$ . The scheduler 7 in the queue management module 5 processes packets from each ingress port in a round-robin manner. For each packet the packet details are transmitted into all the egress queues specified in the destination list for that packet. The time taken for this insertion depends upon the amount of parallelism available  
25 in the queue management module 5, and is referred to as the scheduler bandwidth, which may be 5 insertions per unit time.

Each of the broadcast packets have to be inserted into each of the egress queues (except the source port), so if a broadcast packet arrives in the  
30 ingress queue structure every unit time, the scheduler must have a bandwidth

of  $n-1$  to match the ingress bandwidth (even in the absence of other packets). If the scheduler bandwidth is less than this, the ingress queue sizes will increase.

- 5 Specifically, suppose that the packet rate at each ingress port is  $M$  packets per unit time ( $0 \leq M \leq 1$ ), so that the total number of packets arriving at the switch per unit time is  $NM$ . Suppose that the broadcast traffic as a fraction of all traffic is  $b$  ( $0 \leq b \leq 1$ ), and that the actual scheduler bandwidth is  $S$  per unit time. In this case, the required scheduler rate is  $NM(1-b) + bNM(N-1)$  which
- 10 is equal to  $NM(1 + (N-2)b)$  per unit time. The difference between the egress and ingress rates is thus  $NM(1-b) + bNM(N-1) - S$ , and the rate of increase of the ingress queues is therefore  $\{ NM(1-b) + bNM(N-1) - S \} / N$ .

- In the embodiment, the packet resolution module 1 is arranged to determine
- 15 the length of each of the queues, and according to the lengths determine if BSC should be applied. Preferably, the packet resolution module determines that this is the case when it finds that the length of any one of the queues rises above a predetermined level. Alternatively (or additionally), the packet resolution module may determine that this is the case when it finds that the
- 20 total length of the  $n$  queues (i.e. the sum of the lengths of the  $n$  queues) rises above this predetermined maximum.

- Once BSC has been applied, the packet resolution module 1 continuously monitors whether it must be turned off again. For example, if the BSC was
- 25 triggered by the length of any one of the queues rising above a predetermined level, the BSC may be removed again in the case that it is found that the length of that queue has now fallen below a second predetermined level. Similarly, in the case that BSC was triggered by the total length of the queues rising above the predetermined level, the BSC may be removed in the case

that it is found that the total length of the queues falls below a second predetermined level. In either case, the second predetermined level must be no higher than the first predetermined level, and is preferably lower since this provides a hysteresis.

5

Although only a single embodiment of the method has been described above, the invention is not limited in this respect and many variations are possible, just as there are many known designs of Ethernet switch. In particular, different Ethernet switches manage their ingress ports in different manners,  
10 but the general principle of measuring the lengths of ingress queues and obtaining from them a measure of the proportion of broadcast packets remains valid.